A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

LA-UR--88-4103

DE89 005456

TITLE MILTIGRID MONTE CARLO METHODS

AUTHOR(S) Eugene Loh, Jr.

SUBMITTED TO Proceeding - Computer Simulation Studies in Condensed Matter Physics: Recent Developments Springer Proceedings Physics

DISCLAIMER

MASTER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Multigrid Monte Carlo Methods

Eugene Loh, Jr.
Theoretical Division and Center for Non Linear Studies
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

This paper is intended to be a tutorial on multigrid Monte Carlo techniques, illustrated with two examples. Path-integral quantum Monte Carlo is seen to take only a finite amount of computer time even as the paths are discretized on infinitesimally small scales. A method for eliminating critical slowing down completely — even for models with discrete degrees of freedom, as in Potts models, or discrete excitations, such as isolated vortices in the XY model — is presented.

1. Introduction

Monte Carlo methods have been used quite successfully to investigate many-body systems in condensed-matter physics. Stochastic simulations offer the theorist a tool to examine models in very-high-dimensional spaces in parameter regimes which are inaccessible to analytical methods. Yet many of the interesting phenomena in many-body problems occur at or near critical points in parameter space where the critically slow physical dynamics, which make these points interesting, affect the simulational dynamics as well. Hence, critical phenomena have been a particularly challenging source of difficulties for traditional Monte Carlo techniques.

A variety of attempts have been made to beat such critical slowing down. I would like to mention two other speakers at this workshop in particular. The very first talk was given by Bob Swendsen, who spoke about non-universal dynamics in Monte Carlo simulations. Swendsen and coworkers utilised a mapping between Potts models, such as the Ising model, and percolation models, which do not suffer from critical slowing down, to change the dynamics of Potts simulations. This mapping was first described by FORTUIN and KASTELEYN [1], used for Monte Carlo simulations by SWEENY [2], and adapted differently by SWENDSEN and WANG [3]. The resultant simulation flips disconnected, stochastically defined, clusters at random. Since, at criticality, these clusters are present at all length scales, the dynamics of the simulation slow down much less dramatically with system size than for local-update algorithms.

Also, Alan Sokal described multigrid methods which would block all variables in a local domain together into a single coarse-grid variable. Coarsening deterministically in this way, these simulations would perform lattice updates on every length scale and so, again, change the dynamics of the simulation. Such an approach proves to eliminate critical slowing down completely for "trivial" Hamiltonians such as the gaussian model. For systems with discrete excitations, on the other hand, the acceptance ratio goes down exponentially with the length scale of the Monte Carlo move. Thus, these multigrid methods can only offer savings in computer time by collecting degrees of freedom and processing them together and by thermalizing continuous excitations. The discrete excitations, in contrast, still suffer from critical slowing down. This has been observed in simple-blocking simulations of the XY model above the Kosterlitz-Thouless temperature, where the discrete excitations are the isolated vortices, and in a ϕ^4 scalar-field model

in a double-well potential [4]. Other, self-admittedly disappointing, attempts [5] have been made to generalize the Fortuin-Kasteleyn mapping to arbitrary models. And, of course, there have been many other efforts to address critical slowing down that have not been represented at this workshop.

In this talk, I will try to give a tutorial on multigrid methods by discussing two very simple models: a single quantum-mechanical particle in a harmonic well and the Potts model. From the harmonic oscillator, we will see how high-order interpolation schemes can be used to accelerate simulations of systems with continuous degrees of freedom. The thermodynamics of a quantum particle may be studied by summing over paths of the particle in imaginary time, discretized in nonzero time steps $\Delta\tau$. We will see that the study can achieve the accuracy of an arbitrarily small discretization step even in a finite amount of computer time through the use of multigrid methods. I will also discuss a stochastic coarsening procedure, which was proposed by BRANDT [6] and studied by KANDEL et al.[7]. For the Potts model, the procedure reduces to SWENDSEN and WANG's use of FORTUIN and KASTELEYN's mapping to percolation models. Using the coarsening procedure to employ multigrid ideas provides a means for eliminating critical slowing down completely even in the presence of discrete excitations. This is seen in multigrid simulations on the two-dimensional Ising model.

Most of all, multigrid ideas are only starting to be used in Monte Carlo simulations. What I hope to do, then, is simply to invite others to think about multigridding, using natural geometrical considerations in improving simulational methods, especially in the study of critical phenomena.

2. Quantum mechanical harmonic oscillator

We first consider a single quantum-mechanical particle in a harmonic well. The Hamiltonian describing the motion of the particle is $H = T + V$ with

$$T = \frac{p^2}{2m} \qquad V = \frac{m\omega^2 x^2}{2} \quad , \tag{2.1}$$

where the quantum-mechanical nature of the particle arises from the fact that $[p, x] = -i\hbar \neq 0$. Of course, the natural solution to this model comes from rewriting the position and momentum coordinates in terms of the creation and annihilation operators $b^\dagger$ and b, resulting in the diagonal, second-quantized Hamiltonian $H = \hbar\omega(b^\dagger b + 1/2)$.

Instead, anticipating numerical solutions of less tractable quantum Hamiltonians, we adopt Feynmann's path-integral formalism. While other speakers have already developed this formalism, I present the essential features here once more. Expanding the partition function

$$Z = \mathrm{Tr}\, e^{-\beta H} = \mathrm{Tr}\, e^{-\Delta\tau H} e^{-\Delta\tau H} \cdots e^{-\Delta\tau H} \tag{2.2}$$

as a product of L identical factors, we may now approximate $\exp(-\Delta\tau H) \approx \exp(-\Delta\tau T)\exp(-\Delta\tau V)$ if $\Delta\tau = \beta/L$ is made suitably small. Working in a position-coordinate basis, the factor $\exp(-\Delta\tau V)$ is diagonal and easy to handle. The kinetic-energy factor, $\exp(-\Delta\tau T)$, is simply the free-particle propagator in imaginary time; it is the gaussian

$$< x'|e^{-\Delta\tau T}|x > = \sqrt{\frac{m}{2\pi\Delta\tau\hbar^2}}\ \exp\left(-\Delta\tau\frac{m(x'-x)^2}{2(\Delta\tau\hbar)^2}\right) \quad . \tag{2.3}$$

which, of course, becomes a delta function in the limit $\Delta\tau \to 0$. In this limit, the resultant expression for the partition function becomes a path integral of the exponential $\exp(-S)$ of the action over all paths in imaginary time. For numerical simulations, we will use the approximate expression for Z arising from nonzero

$\Delta\tau$. This so-called Trotter approximation has been well studied [8] and is the foundation of most quantum Monte Carlo work.

People speak of the quantum particle as being represented by a polymer or by a world line in imaginary time. In the latter picture, the world line is described only at discretized times $\Delta\tau$, $2\Delta\tau$, ... β. At each of the discretized times, the world-line coordinate feels an external force from the harmonic potential as well as spring forces from its neighbors in imaginary time. In the limit $\Delta\tau \rightarrow 0$, these spring forces become infinitely strong. Hence, the world line is continuous in imaginary time. Quantum mechanically, however, while such lines are continuous, they are not differentiable. This physical property manifests itself in simulations as statistical noise in measurements of observables which depend on short imaginary-time scales. As one goes to the zero-temperature limit, $\beta \rightarrow \infty$, the "polymer" becomes infinitely long with correlations on the imaginary-time scale $\tau \sim 1/\hbar\omega$. In the classical limit $\hbar \rightarrow 0$, the correlation time diverges and the world lines become completely straight.

Using the Trotter breakup, we have reduced the quantum-mechanical partition function to a high-dimensional sum amenable to classical Monte Carlo techniques. At this point, one may treat the sum using a local-update algorithm [9], accepting or rejecting proposed moves $x(\tau) \rightarrow x(\tau)+\delta$ according to the Metropolis algorithm. Unfortunately, such "wiggles" of the world lines require $\delta \sim \sqrt{\Delta\tau}$ for reasonable acceptance ratios, meaning that Monte Carlo moves must become very small. Furthermore, the Trotter formula assumes $\Delta\tau\ \hbar\omega \ll 1$, meaning that each local move can only affect a segment of the world line which is very short compared to the correlation time. Of course, this situation is exacerbated in the treatment of many-particle systems, for which the correlation times can grow much larger, especially if the system goes critical.

One solution to this difficulty was alluded to by Farid Abraham in his discussion of quantum He on graphite. One may "Fourier accelerate" the simulation by considering nonlocal moves of the form $x(\tau) \rightarrow x(\tau) + \delta_\Omega \cdot \cos(\Omega \cdot (\tau - \tau_0))$ for all points on the world line at once. Not only may one use much larger step sizes δ_Ω for small Ω, accelerating movement through phase space for the long time-scale modes, but one may also sample the various modes with different sampling frequencies. Fourier-accelerated Langevin simulations take advantage of this flexibility by assigning different "masses" to the various fourier modes of the system. For example, if one is interested in measuring only the particle's mean-square displacement, which depends only on $\Omega = 0$ characteristics, then it is straightforward to show that the high-Ω modes should be sampled with frequency $\sim\ \Omega^{-4}$ for optimal statistics. In contrast to most quantum Monte Carlo algorithms, for which more lattice sweeps are needed as $\Delta\tau \rightarrow 0$ to achieve the same quality statistics, this procedure does not slow down since only a finite number of lattice sweeps will be required for high-Ω modes, even as the number of these modes diverges. Put another way, the mean-square displacement depends somehow only on the very longest time scale. The Ω^{-4} rule only gives one a sense of how often the other modes must be sampled to ensure ergodicity — that is, to ensure that the system can sample all of phase space. Of course, the simulation does slow down in the sense that the computer time required for a single sweep increases as $\Delta\tau$ vanishes.

(In contrast to long-time-scale observables, consider measurements of the particle's kinetic energy. Naively, one would measure the expectation value of

$$< x(\tau + \Delta\tau)|T\exp(-\Delta\tau\ T)|x(\tau) > / < x(\tau + \Delta\tau)|\exp(-\Delta\tau\ T)|x(\tau) > \quad , \quad (2.4)$$

giving one the kinetic-energy estimator

$$< T > = \frac{1}{2\ \Delta\tau} - < \frac{m\ (x(\tau + \Delta\tau) - x(\tau))^2}{2\ (\Delta\tau\hbar)^2} > \qquad (2.5)$$

The nondifferentiability of the world line requires the $1/2\Delta\tau$ term for convergence. The estimator (2.5) depends on all modes of the world line and would require one to sample one value of Ω as often as the next. As the imaginary time variable

is discretized on a finer and finer scale, more and more lattice sweeps would be required and yet the measurements would still be more noisy.)

In most Monte Carlo simulations, one updates only one degree of freedom at a time, holding all others fixed for that update. Ergodicity is achieved by subsequently updating other degrees of freedom as well. In our multigrid approach, let us hold fixed interpolations of the world line while updating a position coordinate $x(\tau)$. Consider a short time segment $\tau_0 - \Delta\tau < \tau < \tau_0 + \Delta\tau$ of the world line, which is defined by its coordinates at the discretized times, $x(\tau_0 - \Delta\tau)$, $x(\tau_0)$, and $x(\tau_0 + \Delta\tau)$. We fix a particular interpolation of this segment by fixing the displacement $x(\tau_0) - (x(\tau_0 - \Delta\tau) + x(\tau_0 + \Delta\tau))/2$ of $x(\tau_0)$ from the linearly interpolated position $(x(\tau_0 - \Delta\tau) + x(\tau_0 + \Delta\tau))/2$. Higher- and lower-order interpolations are possible.

To multigrid the simulation, then, we first define the world line of the particle on a very fine grid — that is, the imaginary time is discretized on a very fine time scale. We then "decimate" the position coordinate at every other time step by memorizing its displacement from an interpolated value which is defined in terms of coordinates that are not decimated. Since degrees of freedom have been eliminated from the description of the world line, the cost of updates on the coarser levels decreases inversely with the time scale.

In order to perform updates on the coarser time scale, we must write expressions for the renormalized action of the path. There are two contributions to the action over the segment $\tau_0 - \Delta\tau$, τ_0, $\tau_0 + \Delta\tau$ — one due to the kinetic energy and the other due to the external potential. Writing $x_\pm = x(\tau_0 \pm \Delta\tau)$ and $x_0 = x(\tau_0)$ and fixing $c = x_0 - (x_+ + x_-)/2$, get

$$
\begin{aligned}
S_T &= \Delta\tau \frac{m((x_+ - x_0)^2 + (x_0 - x_-)^2)}{2\,(\Delta\tau\,\hbar)^2} \\
&= \Delta\tau \frac{m((x_+ - (\frac{x_+ + x_-}{2} + c))^2 + ((\frac{x_+ + x_-}{2} + c) - x_-)^2)}{2\,(\Delta\tau\,\hbar)^2} \\
&= (2\,\Delta\tau)\frac{m(x_+ - x_-)^2}{2\,(2\,\Delta\tau\hbar)^2} + \Delta\tau\frac{mc^2}{(\Delta\tau\,\hbar)^2} .
\end{aligned}
\tag{2.6}
$$

Hence the kinetic-energy contribution to the action on the coarser time scale has the same form as that on the finer scale with $\Delta\tau \to 2\Delta\tau$. The second term is simply a contribution to the action which depends only on the particular interpolation, c, and not on the coarse degrees of freedom.

The potential energy contribution to the action is $S_V = \Delta\tau \sum_\tau V(x(\tau))$. The contribution from the segment $\tau_0 - \Delta\tau$, τ_0, $\tau_0 + \Delta\tau$ is therefore

$$
(2\Delta\tau)\,\tilde{V}(x_+, x_-) = \Delta\tau(V(x_+, x_0) + V(x_0, x_-)) \quad , \tag{2.7}
$$

where $\tilde{V}$ is the renormalized potential which depends on the particular interpolation and x_0, of course, is fixed with relation to the interpolated value $(x_+ + x_-)/2$. While $\tilde{V}$ grows in complexity as it is coarsened repeatedly, this is not true in several interesting cases. In particular, for the harmonic potential, V remains quadratic no matter how many times it is coarsened.

The multigrid algorithm, then, is composed of coarsenings, local Monte Carlo updates on the various time scales, and uncoarsenings. To go back to a finer time scale, one simply interpolates between the discretized times and adds back in the fixed displacements. Again, the number of times a fine time scale must be updated decreases rapidly with the time scale — in a quartic fashion $\sim \Omega^{-4}$ for the harmonic oscillator, for example — while the processing cost scales only with the inverse of the time scale. The total cost of processing the very fine time scales, then, is finite even as the finest level is discretized for a smaller and smaller $\Delta\tau$! This is in contrast to fourier accelerated algorithms, whose processing time grows

linearly with $1/\Delta\tau$, and to local-update Monte Carlo, for which processing grows algebraically with $1/\Delta\tau$ even faster than for the fourier moves.

Multigridding is expected to be essential up to the correlation time of the world line — in our harmonic oscillator, this time is $\sim 1/\hbar\omega$. For a system of many quantum particles, this correlation time could be longer and could even diverge critically. Again, the short time scales (Ω large) require only negligibly frequent sampling.

To make the connection to other collective moves, it should be noted that fourier moves correspond to adding cosine waves to the world line. In contrast, GOODMAN and SOKAL's [4] approach may be thought of as adding square pulses. The linear-interpolation scheme mentioned here corresponds to adding triangular pulses whose widths are the time scales at which the pulses are added.

In Fig. 1, the potential energy is plotted as a function of the logarithm G of the number of levels for a single quantum particle of unit mass at inverse temperature $\beta = 5$ in a harmonic well of level spacing $\hbar\omega = 1$. In the limit of few levels, G small, the error due to using a nonzero Trotter parameter $\Delta\tau = \beta \cdot 2^{-G}$ results in large deviations from the $\Delta\tau \to 0$ limit (the dotted line). Three curves are plotted, each representing the same number of processings at the very coarsest time scale, at which measurements of the potential energy were made. While the coarsest level was always sampled the same number of times, finer levels were only visited with relative sampling frequency $\sim \Omega^{-p}$, where, drawing from the fourier picture, Ω is the imaginary-time frequency corresponding to the time scale $\sim 1/\Omega$. The three curves are for $p = 1, 2, 3$. Notice that the three data sets are essentially indistinguishable. Both the data points and the error bars are independent of p; hence, it is not important to visit the fine time scales often. In particular, for $p = 3$, only a finite amount of computer time would be spent on the finest scales even while the number of such levels goes to infinity. Meanwhile, it is also clear from the systematic errors at small G that it is crucial to include these fine time scales to reduce the error due to the Trotter breakup. Notice that this statement is equivalent to the need for ergodicity — including fine time scales is of no consequence if these levels are never sampled. Our rule of thumb from the harmonic oscillator, again, is that $p \approx 4$ offers the best statistics. In our illustration, processing time grows exponentially, 2^G, for $p = 1$, linearly, G, for $p = 2$, and not at all for $p > 2$ as a function of the number, G, of levels.

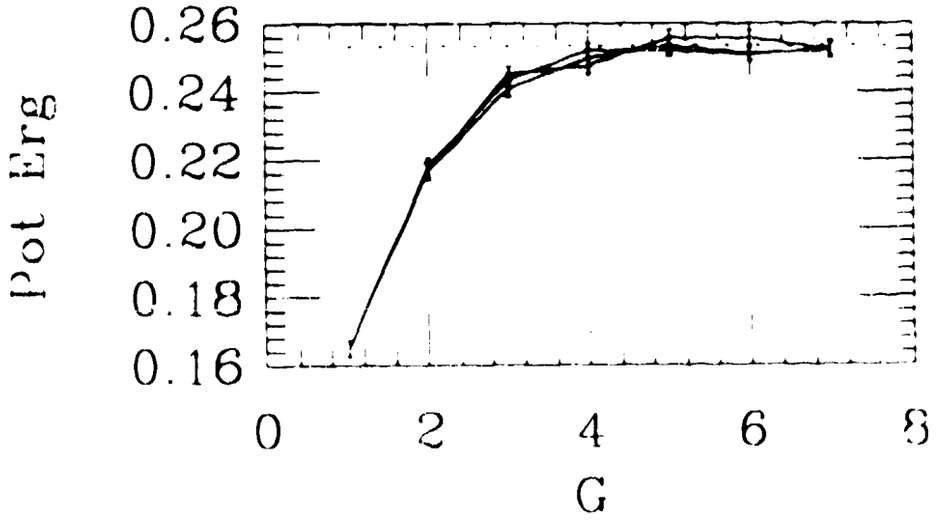

Fig. 1. Potential energy as a function of logarithm of the number of levels for three different sampling frequencies of the short time scales

To close out this section on the quantum particle, we may summarize by saying that multigrid methods allow one to study the limit $\Delta\tau \to 0$ in a finite amount of computer time. (If one chose to store all the path interpolations, however, the memory requirements would grow.) These methods offer efficient numerical solutions for quantum problems — indeed, this single-particle illustration was motivated by work in progress on anharmonic lattice dynamics, whose treatment by these methods is quite straightforward. Multigridding allows one to make large-scale moves through configuration space by stochastically eliminating degrees of freedom on smaller scales.

Nevertheless, one shortcoming, reflective more of path-integral formulations rather than multigrid Monte Carlo itself, must be mentioned. We have discussed measurements of long-time-scale observables, such as the potential energy of the system. Unfortunately, many interesting quantities — the specific heat and kinetic energy, as examples — depend on short-time behavior. As discussed above, however, the nondifferentiability of quantum world lines makes measurement of such quantities difficult. Inclusion of short-time contributions is required to remove systematic errors in the measured averages. Short time scales contribute little to the averages but substantially to the noise. Sampling these fluctuations more often is an unattractive tact as this would be quite time consuming, even in a multigrid approach. Several short-time-dependent observables, fortunately, can be expressed in terms of long-time averages. The specific heat, as many Monte Carlo workers well know, is often best measured by differentiating the energy, which is less dependent on short times than the specific heat itself, than by measuring fluctuations in the energy, which is what a specific-heat operator would do. The kinetic energy, in turn, can be measured using the virial theorem. As a generalization, we note that the expectation value of the commutator of the Hamiltonian $H = T + V$ with any operator A is zero. Thus, $< [T, A] >= - < [V\ A] >$. Since $T \sim p^2$, $[T, A]$ has one fewer factor of x and one more factor of p than does A. On the other hand, V is composed only of position variables and so $[V, A]$ has one more factor of x and one fewer factor of p than does A. This identity, therefore, relates the averages of quantities with different numbers of p's, which have great short-time dependences, and x's, which have only long-time dependence. For example, $A = xp$ gives the virial theorem $< p^2/2m >=< xV'/2 >$. In practice, long-time estimators do not exist for all observables and when they do they are not always well behaved. Alternatively, one can express measurements as sums of contributions from from each of the separate time scales as we saw in (2.6), for example, for the kinetic-energy contribution S_T to the action. The contributions, of course, become progressively more noisy as one goes to shorter time scales. On the other hand, in the limit of short time scales, one may construct estimates of the contributions that depend only on long-time averages. For example, for S_T, the contribution from the shortest time scale is $< m(x_0 - (x_+ + x_-)/2)^2/(\Delta\tau\ \hbar)^2 >$. Since correlations such as $< x(\tau + \Delta\tau)x(\tau) >$ and $< x(\tau + \Delta\tau)x(\tau - \Delta\tau) >$ can be written for small $\Delta\tau$ solely in terms of functions of position coordinates (functions of x, V, and its derivatives), long-time estimates of short-time contributions to averages can be constructed. It is not clear how useful this approach is in practice.

Of course, one could simply live with small $\Delta\tau$ errors and settle for using multigridding only to beat critical slowing down, which may arise in many-body problems.

3. Potts Models

Now let us turn to the question of problems with discrete excitations. In the case of Potts models, for example, changes in the energy clearly must come in finite quanta which are not small on an energy scale set by the critical temperature. I include not only systems described by discrete degrees of freedom, but also models with continuous variables which have discrete excitations. In the classical XY model, for instance, the spin variables are continuous and give rise to continuous excitations such as spin waves. They also allow discrete excitations, however, such as individual vortices, which characterize the phase transition at the Kosterlitz-

Thouless temperature. Our strategy is to design Monte Carlo moves on all length scales with energy changes always of the same, hopefully small, scale. Up to now we have discussed only moves which are everywhere gradual. Unfortunately, for discrete-variable models, such moves are no longer possible. For continuous-variable models, such Monte Carlo moves are possible, but they are ineffectual in thermalizing the discrete excitations. Alternatively, one could consider making non-gradual moves over domains of variables — flipping a prescribed domain of Ising spins, for example, or rotating a selected domain of XY spins. These moves suffer from exponentially decreasing acceptance ratios as the domains grow, meaning that in practice the simulation still does not incorporate large-scale moves.

Here I will describe a stochastic coarsening procedure proposed by BRANDT [6] which allows performing local updates on a coarser length scale even without increasing the scale of energy changes (which would decrease the acceptance ratio). Used in conjunction with multigridding techniques, this procedure allows the simulation of many-body problems without any critical slowing down.

We eliminate the finest length scales and so reduce the number of degrees of freedom with a stochastic coarsening procedure. Disregarding geometrical considerations for the moment, consider a model whose thermodynamics are governed now by a classical Hamiltonian $H = H_0 + V$, where factors of $-\beta = -1/k_B T$ have been absorbed into H. Here, k_B and T are Boltzmann's constant and the temperature, respectively, and H_0 is somehow easier to simulate than the original Hamiltonian. The probability of finding the system in some state Q is proportional to the Boltzmann weight $\exp(H(Q))$, where $H(Q)$ is the energy of the system in state Q. We may "kill" the contribution V to the Hamiltonian stochastically by either "deleting" it with probability $p_d = c_V \exp(-V(Q))$ or by "freezing" it with probability $p_f = 1 - p_d$. If the interaction is frozen, only states Q' with $V(Q') = V(Q)$ are considered in the ensuing simulation. If the interaction is deleted, no such restriction is placed on the states. In either case, the thermodynamics are subsequently governed only by the simplified Hamiltonian H_0. The coefficient c_V must be chosen so that $p_d, p_f \in [0, 1]$ — p_d and p_f must be probabilities. The largest choice of c_V produces the best statistics. By assumption, H_0 is easier to study than H and so the simulation will proceed more efficiently than before.

Clearly, this procedure is strongly ergodic since there is always a nonzero probability that no restriction will be placed on the simulation, allowing nonzero transition probabilities between all states. It also satisfies detailed balance. To see this, first consider two states Q and Q' with $V(Q) \neq V(Q')$. Then, a transition from one state to the other can take place only if V has been deleted:

$$T(Q \longrightarrow Q') = c_V \, e^{-V(Q)} \cdot \frac{e^{H_0(Q')}}{Z_0} \quad , \tag{3.1}$$

where Z_0 is the partition function for the reduced Hamiltonian. Now,

$$\frac{T(Q \longrightarrow Q')}{T(Q' \longrightarrow Q)} = \frac{e^{-V(Q)} \, e^{H_0(Q')}}{e^{-V(Q')} \, e^{H_0(Q)}} = \frac{e^{H_0(Q')+V(Q')}}{e^{H_0(Q)+V(Q)}} = \frac{e^{H(Q')}}{e^{H(Q)}} \quad . \tag{3.2}$$

Alternatively, if $V(Q) = V(Q')$, the interaction V may either be deleted or frozen:

$$\begin{aligned} T(Q \longrightarrow Q') &= c_V \, e^{-V(Q)} \cdot \frac{e^{H_0(Q')}}{Z_0} + (1 - c_V \, e^{-V(Q)}) \cdot \frac{e^{H_0(Q')}}{Z_0'} \\ &= \left(\frac{c_V \, e^{-V(Q)}}{Z_0} + \frac{1 - c_V \, e^{-V(Q)}}{Z_0'} \right) e^{H_0(Q')} \quad , \end{aligned} \tag{3.3}$$

where Z_0' is the partition function for the reduced Hamiltonian over the restricted space. Then, using $V(Q') = V(Q)$, we find

$$\frac{T(Q \longrightarrow Q')}{T(Q' \longrightarrow Q)} = \frac{e^{H_0(Q')}}{e^{H_0(Q)}} = \frac{e^{H_0(Q')+V(Q)}}{e^{H_0(Q)+V(Q)}} = \frac{e^{H_0(Q')+V(Q')}}{e^{H_0(Q)+V(Q)}} = \frac{e^{H(Q')}}{e^{H(Q)}} \quad . \tag{3.4}$$

completing the proof of detailed balance.

In practice, H_0 is still nontrivial to simulate efficiently and so additional terms of the Hamiltonian must be killed. After killing all the interactions in H, one arrives at a system which is completely decoupled — and so is trivial to simulate — but it is subject to an arbitrary set of restrictions on its states.

As an example, consider the Ising model $H = \sum_{<i,j>} K_{ij} s_i s_j$. The optimal probability for deletion is $p_d = \exp(-K_{ij}(1 + s_i s_j))$. Interactions between antiparallel spins will always be deleted; only parallel spins can be frozen together in ferromagnetic models. We may kill the interactions $K_{ij} s_i s_j$ one at a time until we are left with irregular, fractal blocks of spins which are completely decoupled from one another. This coarsened system is trivial to study and so statistics over many such coarsenings are easily gathered. This procedure differs from standard block-spin projection methods in that here blocks are generated in a stochastic manner and are generally of irregular shape. As one can see, for Potts models the coarsening procedure is identical to that used by SWENDSEN and WANG [3]. Due to the restrictions that are introduced, however, Swendsen and Wang still observe critical slowing down, albeit with a considerably reduced dynamical exponent.

To eliminate critical slowing down completely, we incorporate the stochastic blocking procedure as part of a multigrid scheme. Instead of coarsening the system until all that remains are large, decoupled blocks of spins, we coarsen out only the very finest length scales at each level of the multigrid algorithm. The coarsened system is composed of small blocks, typically all of some an all length scale b, which interact according to a reduced Hamiltonian. We have explored several ways of killing only fine length scales; perhaps the most transparent, though perhaps also not the best, is one in which some fraction of the bonds selected at random are killed. The coupling between two coarse blocks is the sum of the living couplings that connect fine-lattice spins frozen to those blocks. Notice that long-range interactions may be generated, but they are improbable and their presence does not influence the convergence of the algorithm. The resultant "simplified" system is stored. It is studied by further coarsening.

While Swendsen and Wang coarsened their Potts lattices completely for each iteration of their simulations, we return to intermediate length scales. In particular, we coarsen the lattice γ times at each level of processing before returning to the next finer level. In this language, Swendsen and Wang, in effect, use $\gamma = 1$. To "uncoarsen" the system, decoupled blocks should be set to some arbitrary value of spin, all fine-lattice spins should be set to the block spin to which they were frozen, and the fine-lattice couplings should be restored. Metropolis updates may be performed at any length scale by using a standard Metropolis algorithm on the block spins at that length scale using the Hamiltonian appropriate to those blocks.

Our Monte Carlo method "cycles" through all the various length scales.[4] At each intermediate length scale, the system is coarsened γ times before it is uncoarsened. Each time the system reaches the coarsest level, at which all the blocks are decoupled, measurements may be made and the system is immediately uncoarsened. The cycle ends each time the finest level is reached. A few Metropolis sweeps are performed between coarsenings and uncoarsenings. These Metropolis sweeps are not essential to defeating critical slowing down and, in practice, performing more than one such sweep at a time is ineffective in accelerating the procedure.

We carried out simulations of the $d = 2$ Ising model on square lattices from 4^2 to 128^2 sites with periodic boundary conditions, using a cycle of $\gamma = 2$ with rescaling factor $b = 2$. Starting from fully magnetized states, we measured the decay of the energy to its equilibrium value for an ensemble of configurations. The energy relaxation was described by an exponential decay with, surprisingly, no discernable short-time transients. The relaxation times were $\tau(L) = 3.6 \pm 0.5$, independent of linear size L for $L \geq 16$. In contrast, other Monte Carlo algorithms show critical slowing down — $\tau(L) \sim L^z$ — with dynamical exponents $z \approx 2.1$ for standard single-spin-flip Monte Carlo and $z \approx 0.35$ for Swendsen and Wang's

method. Similar measurements for the magnetic susceptibility at criticality showed the relaxation time saturating at $\tau = 7 \pm 2$, again at $L = 16$, for our algorithm.

But how does our approach eliminate critical slowing down? The first answer is simply that it does and not every algorithm that allows large-scale Monte Carlo moves achieves this. If the coarsening procedure tended to create coarse lattices with higher connectivities or stronger bonds than those of the fine lattices, the acceptance ratios for the large-scale moves would become prohibitively small. On the other hand, if the connectivities tended to be lower or the couplings weaker, blocks would become decoupled at short length scales and large-scale flips would not be possible. Thus, in addition to the empirical evidence that our multigrid Monte Carlo algorithm produces a dynamical exponent $z = 0$, it is important to note that we find that our coarsening procedure yields similar distributions of bonds and connectivities in lattices at different length scales.

Swendsen and Wang also produce Monte Carlo moves that flip blocks at all length scales. Why then do they not achieve $z = 0$? Consider, again, the multigrid coarsening procedure, which essentially treats interactions first at the finest length scales and then at increasingly longer length scales. The reason this is only "essentially" true is that the various length scales cannot be treated completely independently — to some extent, whenever an interaction is "frozen" at a fine length scale, restrictions on the allowed states are introduced at all coarser length scales as well. For example, if interactions are killed (more importantly, frozen) up to some length scale, then the coarsened Hamiltonian were simulated infinitely fast, and finally the fine-lattice spins restored and the procedure iterated, there would still be a correlation time. That time would grow with the number of levels that had been frozen and from scaling arguments we would conclude that the rate of growth would be independent of level number. Physically, it would be nice to decouple all the various length scales. In practice, we do not know how to design stochastic coarsening procedures that will do this for nontrivial models. Results on $\gamma = 1$ cycles tell us to what extent a particular coarsening procedure interlocks the length scales, measuring the amount of correlations that have built into the simulation by killing interactions over some number of length scales. Thus, each level must be visited $\gamma_{\min}$ times more often than the next finer level, where $\gamma_{\min} = b^{z_1}$, b is the length rescaling factor which describes the degree of coarsening that takes place between consecutive levels, and z_1 is the dynamical exponent for $\gamma = 1$ cycles. We find that z_1 is in fact scale invariant and, for the coarsening procedure that we and Swendsen and Wang use, is $z_1 = 0.35$ for the $d = 2$ Ising model.

Exponentially many coarsenings must take place at the coarser levels. This increased processing is acceptable since processing at coarser levels is cheaper due to the smaller number of degrees of freedom. Indeed, the work function is still proportional only to lattice size so long as the increase in the amount of processing is smaller than the amount by which the number of degrees of freedom is reduced: $\gamma < b^d$, where d is the dimensionality of the system. This is not a practical difficulty since the dynamical exponent for the $\gamma = 1$ cycle is typically much smaller than the dimensionality of the system: for the Ising ($q = 2$ Potts) model, complete elimination of critical slowing down for a finite amount of processing time per site requires $b^{0.35} < \gamma < b^2$ in two dimensions and $b^{0.75} < \gamma < b^3$ in three dimensions. For the 3-state Potts model in two dimensions, $b^{0.6} < \gamma < b^2$. In our simulations for the two-dimensional Ising model, we chose $\gamma = 2$ and $b = 2$, which is clearly within the regime of no critical slowing down.

Aside from geometrical considerations, it is clear that it is most efficient to perform most of the processing at the coarsest levels, for which there are the fewest degrees of freedom. The role of multigrid ideas and the above scaling arguments is to suggest coarsening schemes and parameter regimes for which the processing time is not only lowered but, in fact, critical slowing down is eliminated completely.

4. Conclusions

In multigrid methods, one coarsens the degrees of freedom to reduce the computational complexity of the problem and to enable local processing at coarse levels to effect large-scale changes on the fine lattice. In multigrid Monte Carlo, random processes are used both to represent coarsened interactions stochastically, but exactly, and also to update the coarsened degrees of freedom with local Monte Carlo moves.

In the longer range, the application of multigrid Monte Carlo to frustrated systems would be interesting. Uniform frustration presents no special difficulties. In contrast, it is not clear how best to model spin glasses, which have random frustration.

Some of this work was motivated by studies of dynamics. While the advantage of techniques that beat critical slowing down is that they change the dynamics of the system, relating the dynamics of a d-dimensional model to the equilibrium statistical mechanics of a $d+1$-dimensional problem [10] allows one to employ multigridding for nonequilibrium studies.[11]

Finally, a number of interesting technical questions remain. How, for instance, does one optimize the coarsening procedures to reduce statistical noise and to minimize the number of visits to the finest levels? What ways are there of constructing estimators which have minimum noise and depend only on very large scales? How can one further simplify complicated effective potentials on coarse length scales stochastically? And what detailed tests can be performed of the scaling conjectures presented above?

But most of all, multigrid Monte Carlo is still a very young class of techniques. Clearly, a great deal of work in the near future will consist simply of knocking down straw men — multigridding straight-forward Monte Carlo simulations in order to draw up a more lengthy list of credentials. Among such systems are models with both continuous and discrete excitations, such as the XY model, with its spin waves and independent vortices, and many-boson Hamiltonians, which are described by continuous world lines as well as discrete exchange effects.

1. P.W. Kasteleyn and C.M. Fortuin, J. Phys. Soc. Jpn. Suppl. **26s**, 11 (1969); and C.M. Fortuin and P.W. Kasteleyn, Physica (Utrecht) 57, 536 (1972).

2. M. Sweeny, Phys. Rev. **B27**, 4445 (1983).

3. R.H. Swendsen and J.S. Wang, Phys. Rev. Lett. **58**, 86 (1987).

4. J. Goodman and A.D. Sokal, Phys. Rev. Lett. **56**, 1015 (1986).

5. R.G. Edwards and A.D. Sokal, unpublished.

6. A. Brandt, Multilevel Computations: Review and Recent Developments, in Preliminary Proceedings of the Third Copper Mountain Conference on Multigrid Methods (April 1987).

7. D. Kandel, F. Domany, D. Ron, A. Brandt, and E. Loh, Jr., submitted to Physical Review Letters.

8. See, for example, M. Suzuki, J. of Math. Phys. **26**, 601 (1985).

9. M. Creutz and B. Freedman, unpublished.

10. See, for example, E. Domany, Phys. Rev. Lett. **52**, 871 (1984).

11. D. Kandel, to be published.